DOCUMENT RESUME

ED 083 284 TM 003 246

AUTHOR	Klein, Stephen P.; Kosecoff, Jacqueline
TITLE	Issues and Procedures in the Development of Criterion Referenced Tests.
INSTITUTION	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	TM-R-26
PUB DATE	Sep 73
NOTE	18p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Criterion Referenced Tests; Elementary Grades; *Mathematics; Secondary Grades; *Test Construction; *Testing Programs; *Tests

ABSTRACT

The basic steps and procedures in the development of criterion referenced tests (CRT), as well as the issues and problems associated with these activities are discussed. In the first section of the paper, the discussions focus upon the purpose and defining characteristics of CRTs, item construction and selection, improving item quality, content validity, item and test bias, test scores, and packaging and other considerations. In the second section, the results of a survey conducted to assess current efforts in criterion referenced testing are summarized. Five defining characteristics--program focus, instructional dependence, objective and item generation, test models and packaging, and test scores--are provided for each of the following testing programs: California Test Bureau--McGraw-Hill, Prescriptive Mathematics Inventory; Comprehensive Achievement Monitoring; Individualized Criterion Referenced Testing; Instructional Objectives Exchange; MINNEMAST Curriculum Project--University of Minnesota; National Assessment of Educational Progress; Southwest Regional Laboratory; System for Objectives Based Assessment--Reading, Center for the Study of Evaluation; UCLA; and Zweig and Associates. From this analysis, 10 questions that the CRT developer must answer in order to clarify the nature and purpose of a CRT are provided. (DB)

FILMED FROM BEST AVAILABLE COPY

ED 083284

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

Conducted by Educational Testing Service in Association with Rutgers University Graduate School of Education

TM REPORT 26 SEPTEMBER 1973

ISSUES AND PROCEDURES IN THE DEVELOPMENT OF CRITERION REFERENCED TESTS

Stephen P. Klein
Jacqueline Kosecoff

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 003 246

PREFACE

A visitor to our planet Earth surveying the current state of educational testing would very likely be confused by what he found. He would observe, for example, the increasing use of tests in all phases and facets of the educational process including the evaluation of instructional personnel. He would learn, too, about the great technological improvements that have been made in tests and in their administration, scoring, and reporting procedures. All of these factors would tend to support the notion that tests are fulfilling an important and vital role. On the other hand, this same observer might also hear the valid complaints of the growing cadre of test critics. These critics complain that present tests are inappropriate for most educational decision making and, if a test is not going to be used for decision making, why bother giving it in the first place?

Perhaps one of the quickest ways of alleviating our visitor's confusion is to point out to him certain changes that have been occurring in education and testing during the past few years. For example, most expert test construction in the past has focused upon a relatively few kinds of assessment instruments, such as those that are used to decide whether a student should be accepted for college. Comparatively little help has been given to the classroom teacher to diagnose individual student needs or assess the outcomes of particular instructional programs. Now, however, there is growing desire to individualize instruction, to assess validly the outcomes of instructional programs, and to hold teachers and administrators responsible for actual gains in student performance. These trends have increased the demand on test developers for appropriate tools to facilitate the measurement process, because existing measures are useful for some important educational decisions but are not designed to meet all needs. It is evident, therefore, that test critics are complaining not about tests per se, but about the need for certain kinds of quality measures that are not currently available.

It is within this context of increased need for and reliance on valid test results that the movement towards so-called "criterion referenced tests" (CRT) has been given new impetus. A criterion referenced measure is essentially "one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards."[1] (Glaser & Nitko, 1971). The pertinent question is whether or not the individual has attained some significant degree of competence on an instructional performance task (Harris, 1972).

Measures with these characteristics are, of course, not new to education. What is new is the range of importance of the decision areas for which they are being employed or emphasized and the attention they are being given by measurement and curriculum experts alike (Airasian & Madaus, 1972; Baker, E., 1972; Keller, 1972; Davis, 1972, 1973; Hawes, 1973). It would not be surprising, therefore, for us to witness during the next few years a number of major contributions to testing theory and methodology arising from the use of criterion referenced tests. Further, the improvement of such measures is likely to have many ramifications for instructional practice, since with improved tools even more reliance is likely to be placed upon the results obtained. For example, a bill is now pending before the United States Congress that would require criterion referenced test data in order to make funding decisions

[1] These performance standards are usually behaviorally stated, for example: "The student will be able to perform all fundamental mathematical operations involving single-digit integers."

This publication was prepared pursuant to a contract with the National Institute of Education, U. S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy.

affecting thousands of schools and involving several million dollars (Quie, 1973).

It is appropriate at this point in time, therefore, for us to examine how criterion referenced tests are constructed and, more importantly, the basic issues and procedures associated with these steps. It is hoped that such an appraisal will clarify some of the basic methodological and theoretical concerns associated with criterion referenced tests that will be examined during the ext few years.

This paper is divided into two parts. In the first section the major issues and steps in the development of CRTs are considered. In the second section representative CRT systems in mathematics, as well as important efforts in other content areas, have been selected for review.[2]

[2]The intended focus of this paper was to be CRTs in mathematics; however, a review of the relevant literature disclosed relatively few references dealing exclusively with this field. Further, those articles pertaining specifically to mathematics mainly describe the development of particular instruments in certain contexts. They do not consider the general concerns associated with the development and use of CRTs in mathematics nor do they examine the vast array of situations for which they might be applicable. Therefore, it was decided to focus this paper on concerns central to CRTs in general, with special emphasis and examples coming from mathematics.

MAJOR ISSUES AND STEPS IN CRT DEVELOPMENT

This section of the paper provides a review of the basic steps in the development of CRTs and the major issues associated with these steps. Although many of the steps and issues have their counterpart in classical test development, the present focus is upon those considerations unique to CRTs and especially those relating to the development of such measures in mathematics. It should be kept in mind, however, that the method chosen to resolve a particular issue at one stage in the development of a CRT is likely to have ramifications for other stages in the developmental process as well as in the interpretation of the scores obtained. In addition, the most important but not necessarily self-evident of these implications are noted, and the primary techniques and procedures that have been used as well as their most important advantages and limitations are identified.

Purpose and Defining Characteristics of CRTs

It is a generally accepted principle that somewhat different kinds of measures have to be constructed for different purposes. This principle also appears to carry over into the development of CRTs. For example, to ensure an adequate level of test reliability, a CRT or series of CRTs that will be used in making a decision about an individual's level of performance will need to be longer than one used for group assessment. Similarly, the focus of the CRTs used for managing an individualized learning segment of a small mathematics unit would be narrower than that used to measure end-of-year performance of all students in a classroom. The characteristics of the target audience, such as their ages and ethnic backgrounds, are also likely to influence the test construction process in terms of the appropriateness of various kinds of stimuli and response formats. Further, the anticipated number of students to be tested and the context in which the testing will occur influence test format, production, distribution, administration, scoring, and analysis.

Figure 1 lists some of the basic purposes that have been noted for using CRTs in terms of the decisions to be made and the focus of the testing (Harris, 1973; Skager, 1973). Three major kinds of decisions have been identified. Decisions relating to the organization of an instructional program are classified as planning decisions. Validating the quality and competency of a program is encompassed by certification decisions. Decisions based on additional investigation of the instructional program are included in a research category. With respect to the focus of the testing program, three classifications are considered. First, a CRT can be primarily involved with the individual student. Second, groups of students such as a classroom or ethnic group can be the focus. And third, the instructional program itself might be the primary unit of concern.

Figure 2 illustrates how differences in the target audience would result in different test items for the same objective. From an inspection of these figures and the foregoing discussion it is apparent that the different uses of CRTs may require different kinds of measures and test models. The fundamental issue underlying these differences is the degree to which the CRT or set of CRTs will provide precise and reliable information about student performance relative to various feasibility constraints associated with gathering this information, such as costs and testing time.

Figure 1. Purposes for Criterion Referenced Tests

FOCUS OF THE TESTING PROGRAM	*TYPE OF DECISION* PLANNING	CERTIFICATION	RESEARCH
Student	Diagnosis, Prediction, and Placement	Determination of "mastery," grades, and success of placement	Interactions between the student, the group, and the program
"Group" (Classroom, ethnic, SES cultural, or geographical groups)	Classroom management Curriculum selection	Instructional and administrative accountability	Interactions between group(s) and program, e.g., do students with certain characteristics function better than others in a given situation?
"Program" (A program may be used with one or more groups)	Organization and sequencing of instruction, Curriculum and product development, Needs Assessment	Program Evaluation Analysis of subject matter domain	Comparisions between types of programs Analysis of program components Development of measurement methodology

Figure 2. Comparison of General Item Formats for the Same Objective at Different Grade or Age Levels

OBJECTIVE

The student will indicate by marking the appropriate choices on an attitude scale his/her appreciation of the importance of mathematics in everyday life.

	FIRST GRADERS	TWELFTH GRADERS
Format.	The student is given a test booklet. Each page is a different color and has a familiar symbol at the top of the page, such as a rabbit. Each page also has the words "Yes" and "No." Directions are provided to the student so that he/she understands to mark the choice that answers the question that is read by the teacher.	The student is given a set of statements and a series of choices ranging from "Strongly Agree" to "Strongly Disagree." The student marks the number of his choice on a machine-scorable answer sheet.
Sample Items	The teacher reads the following kinds of directions and questions: "Now turn to the red page with the rabbit at the top. . . . Now I am going to read you the next question. 'Do you have to know how to work with numbers to tell time?'. . . . Now turn to the yellow page with the duck at the top. . . . 'Do you have to know how to add and subtract numbers to catch a ball?'. . . . Now turn to the page with the table at the top. . . . 'Do you have to know how to work with numbers to buy something at the store?'. . . . and so forth.	The following kinds of items might appear on a scale to measure the objective: 1. Persons who fill medical prescriptions need to use mathematics frequently in their work. 2. Only a very small part of a carpenter's job requires him to use mathematics. 3. It is more important for a bank teller to make friends easily than it is for him or her to make arithmetic computations accurately. 4. In order to be a good plumber, one would have to be able to do basic arithmetic computations with fractions.
Comments.	Note that the child does not have to read the questions, the questions are asked about him or herself rather than some other person, and that the language level and activities are within the students' repertoire of experiences.	The statements are balanced with respect to being positive or negative regarding the importance of mathematics so as to reduce any irrelevant tendency to agree or disagree.

Objectives Chosen

As noted in the preface to this paper, one of the essential features of CRTs is their foundation in clearly defined educational objectives. There are, however, a number of issues associated with how these objectives should be developed and stated. The essence of these issues may be summarized by the question: "What kinds of objectives should form the basis for a CRT system?"

Almost all developers of CRTs agree that to assess performance within a given area requires the construction of a set of CRTs rather than a single measure. The problem then arises as to which objectives within an area should become the basis for the CRTs and how broadly or narrowly these objectives should be stated, that is, the extent of each *objective's coverage*. The statement of an objective may be further delineated by defining the *conditions* under which the measurements are made (e.g., open vs. closed book, with or without the aid of a sheet containing needed formulas, and so forth) and/or the *standards* of performance to be reached in order for the objective to be achieved (e.g., "80 percent correct ," "in less than 2 minutes," and so forth) (Mager, 1962; Popham, 1965). Implicit or explicit assumptions about the *relative importance* of the objectives and the *characteristics of the area* to be assessed (such as the logical and/or sequential organization of the objectives in it) also influence decisions as to which objectives should form the basis for a CRT system (Popham, 1972).

The resolution of the issues associated with choosing a set of objectives usually hinges upon the anticipated purpose(s) of the CRT system. Thus, there is a consideration of the degree of precision needed relative to various practical considerations. This balance is illustrated by the IOX Criteria for Objective Selection (Popham, 1972) presented in the Appendix.

Some of the procedures that have been used to develop the objectives bases for CRTs systems are described briefly below:

1. Expert Judgment. A small group of experts within the area to be assessed meet and, on the basis of their knowledge and experience in the field, jointly decide which objectives are the most important to measure. These objectives are then screened to determine the feasibility of measuring them and, where necessary, to clarify and/or redefine them. This is probably the most common approach.

2. Consensus Judgment. Various groups such as community representatives, curriculum experts, teachers, and school administrators decide which objectives they consider to be the most important. A measurement and/or curriculum expert is then responsible for defining and stating these objectives in a way that would permit them to be assessed (Klein, 1972; Wilson, 1973).

3. Curriculum Analysis. A team of curriculum experts analyzes a given set of curriculum materials such as textbooks in order to identify, and where necessary infer, the objectives that are the focus of these materials (Baker, R.L., 1972).

4. Analysis of the Area to be Tested. An in-depth analysis is made of an area such as mathematics in order to identify all contents (such as single-digit numerals) and behaviors (such as multiplication with replacement) that are included in that area (Glaser & Nitko, 1971; Nitko, 1973). The objectives associated with these contents and behaviors are then organized in some systematic fashion, such as in terms of a hierarchy and/or sequence of objectives for the components of the subject area (in mathematics usually referred to as "strands") (Nitko, 1971; Roudabush, 1971; Popham, 1972).

Item Construction and Selection

Once the purpose(s) and the objectives for the CRT system have been delineated, the next step is to construct and/or select test items or tasks to measure the objectives chosen. This is one of the most difficult steps in the total developmental process because of the vast number of test items or tasks that might be constructed for any given objective, even those that are relatively narrowly defined. For example, consider the following objective: "The student can compute the correct product of two single digit numerals greater than 0 where the maximum value of this product does not exceed 20." The specificity of this objective is quite deceptive since there are 29 pairs of numerals that meet this requirement and at least 10 different item types that might be used to assess student performance (see Figure 3). Further, each of the resulting 290 combinations of pairs and item types could be modified in a variety of ways that might influence whether the student answered them correctly. Some of these modifications are:

- Vary the sequence of numerals (e.g., 5 then 3 versus 3 then 5).
- Use different item formats (e.g., multiple choice versus completion).
- Change the mode of presentation (e.g., written versus oral).
- Change the mode of response (e.g., written versus oral).

It soon becomes evident that even a highly specific objective could have a potential item pool of well over several thousand items (Hively, 1970, 1973; Bormuth, 1970).

The number of items to construct for each objective is influenced by several factors. Some of these factors are the amount of testing time available and the cost of making an

Figure 3. Item types using the content of numerals 3 and 5 for the objective

The student can compute the correct product of two single digit numerals greater than 0 where the maximum value of this product does not exceed 20.

a. $\begin{array}{r} 5 \\ \times 3 \\ \hline \end{array}$

b. 5 x 3 =

c. (5) (3) =

d. 5 · 3 =

e. 5 times 3 =

f. The product of 5 and 3 =

g. 5 x ___ = 15

h. If x=5 and y=3, what is the value of xy?

i. What numeral multipled by 3 will equal 15?

j. John has 5 apples. Sally has 3 times as many apples as John. How many apples does Sally have?

interpretation error, such as saying that a student has achieved mastery when he has not. A survey of current measures reveals that the usual practice is to use about three to five items per objective. This practice appears to stem more from feasibility constraints than any sound foundation in psychometric theory or technology.

The particular item construction and selection approach or combination of approaches chosen to define a CRT program is a major consideration. One reason for this is that the methods used have a direct bearing on the utility and content validity of the CRTs developed and the interpretation of their scores. For example, if there is a hierarchy of objectives and if a CRT is to be based on an objective at a given level of generality in this hierarchy, then it is likely that the items used will be sampled from the relevant subobjectives. Unless there is a specified hierarchy or an organization of objectives, such systematic sampling is impossible. When this latter situation occurs, one has much less confidence that the measure(s) developed really assess the whole objective. One reason for this concern is that without a systematic plan for guidance, it is very easy to just construct items for those aspects of an objective that are most amenable to measurement rather than those aspects that might be considered most germane or critical. On the other hand, it also seems likely that responsible test developers working without an overall plan are more likely to focus their attention on the most salient (and perhaps most frequently taught) facets of an objective than on those aspects that may be just tangential to what a student must really know or be able to do. Thus, the best compromise between systematic sampling (and thereby improved content validity) and potential instructional relevance is to first develop a provisional systematic plan and then assign items to some or all the components of this plan based upon their perceived relative importance. This latter approach is the one most frequently adopted by major test publishers (Wood, 1961).

A related issue in construction and selection of CRT items is the degree to which the items should be sampled with respect to their relative difficulty within an objective. It is a well known and frequently used principle of test construction that slight changes in an item can affect its difficulty. This is most readily accomplished by varying the homogeneity of the alternatives in a multiple choice item, such as in the two examples below:

Eight hundreths equals	Eight hundredths equals
a. 800	a. 800
b. 80	b. .80
c. 8	c. .08
d. .08	d. .008

The extent to which the items within an objective are sampled with respect to difficulty has, of course, a direct bearing on the interpretation of the scores obtained. In other words, if only the most difficult items are used, then the phrase "mastery of the objective" has a very different meaning than if the items were sampled over the full range of difficulties. The fact that the difficulties of items on CRTs (and thus their scores) can be influenced so easily poses real problems to CRT users. To blindly assume that the scores obtained indicate an accurate appraisal of the degree of mastery achieved, merely because a measure is called a "CRT" is an exercise in self-deception.

A third consideration influencing the construction and selection of items is the degree to which an item is dependent upon or related to a particular set of curriculum or institutional materials and techniques (Baker, R., 1972; Skager, 1973). For example, if the instruction only gave students practice in solving multiplication problems in the form used in item types a-e in Figure 3, and if the CRT for this unit only used these same item types, then the CRT would be said to be "instructionally dependent" or biased. It is readily apparent that the more instructionally dependent the CRT, the more likely the effects of instruction would be evidenced in the scores obtained with it and the less generality one could draw from these scores regarding the student's mastery of the objective. On the other hand, instructionally independent tests are more likely to reflect a student's general ability. Thus, an instructionally biased test might be preferred for such purposes as teacher accountability, while an instructionally

independent test might be preferred for school accountability and for evaluation studies comparing the effects of different programs.

A fourth issue, and one which has perhaps not received as much attention as it should, is the potential interaction between the objective and how it is measured. It is often assumed, for example, that selected response items (e.g., multiple choice) serve as an effective proxy for constructed response items (e.g., completion or short answer) because the performance of students on the two kinds of items are highly related. Although this may be generally true, it may not be true for certain kinds of objectives; and further, the degree of mastery required to answer a constructed response item is usually greater than it is to answer the selected response item. The relative scoreability of the latter format, however, has led to its use almost exclusively in published measures, including CRTs. It should be recalled that anything affecting item difficulty on a CRT will influence the total score on it and thereby the interpretation of that score.

The foregoing considerations have led to a number of different methods of selecting and constructing items for CRTs. The general features of these methods are described below, but it should be remembered that each of these approaches begins with or involves the development of well-defined statements of the educational objective(s) to be measured.

1. Panel of Experts. A group of measurement and curriculum "experts" decide which items to use based on their knowledge and experience of the field (Zweig, 1973). When the experts involved are classroom teachers, this approach may lead to highly instructionally dependent measures.

2. Systematic Sampling. This approach is basically a variation of the classical test construction technique. It involves developing for each objective a matrix of contents and behaviors (or tasks) to be assessed. Items are then systematically sampled within this matrix and perhaps along a third continuum of item difficulty as well (Wilson, 1973; CTB/McGraw-Hill, 1973).

3. Systematic Item Generation. This is the most sophisticated of the various approaches and starts with the assumption that all the relevant contents, behaviors (or tasks), stimulus and response characteristics, and related factors can be defined for a given domain or universe of objectives (Hively, 1970, 1973; Cronbach, 1971; Skager, 1973). Basic item forms or "shells" are then constructed. Various techniques can then be used to generate the necessary items, including the use of a computer (especially in the field of mathematics) to meet certain prespecified criteria for coverage of the objectives (Kriewall & Hirsch, 1969).

It is evident from these descriptions that as the sophistication of the method improves, generality of the results and the costs of test construction tend to increase. Further, the particular method chosen will be influenced by the nature of the efforts that have been devoted to the generation of the objectives on which the CRTs are based and the purposes for which they will be used. Finally, the degree of sophistication may be limited still further by the clarity of the domain to be assessed, such as mathematics versus "citizenship" and the measurement technology available for constructing measures in that domain (e.g., academic achievement versus personality development).

Improving Item Quality

It is axiomatic that all tests and measures be field tested prior to basing decisions upon them. Although it appears that this axiom is often ignored, there are a number of methods that have been suggested for analyzing CRT items in order to identify those that are "faulty." It should be noted, however, that an item that is considered "faulty" or "good" using one method of analysis may not be identified as such using another method (Popham & Husek, 1969). This is illustrated in Table 1. It is apparent, therefore, that the final version of a test may be influenced greatly by the method of item analysis chosen for its construction (Cox & Vargus, 1966; Roudabush, 1973).

Table 1. Results of Different Item Analyses

Item No.	Item Difficulty		Possible point biserial r with score on test	Possible sensitivity to instruction
	Pretest	Posttest		
1	0%	100%	0	High
2	50%	50%	1.00	Low

There are two basic concepts underlying the item analysis techniques associated with CRTs and at least one of these constructs is present in each analysis method. These two constructs are as follows:

1. An item is considered "good" if it is *sensitive* to instruction, that is, if performance on it is related to the degree of instruction obtained. The methods that rely heavily on this construct are usually used when there is little or no variation in student scores at any one testing. There are problems with such methods, however: they assume that the instruction was indeed effective; they tend to produce instructionally dependent measures; and they are biased by maturation and other irrelevant systematic factors that might tend to improve scores over time. Further, the use of a technique emphasizing sensitivity could easily lead to

some rather interesting circular reasoning if one tried to improve the test and an instructional program at the same time.

2. An item is considered "good" if it *discriminates* between those who did well versus those who did poorly on the test as a whole or some "outside" criterion, such as performance in the next step in a sequence of instruction. This involves all the classical item analysis approaches and as such one must accept all the assumptions, advantages, and disadvantages that are normally associated with these techniques (especially item and test variance).

The kinds of analysis methods and their variations that have been suggested are listed below:

1. Comparison Group. Give the test to two groups who are known to possess different degrees of skill with respect to the objective(s) measured. One way of doing this is to give the test to those who have versus those who have not received instruction dealing with the objective. A second method is to give the test to those whose normal activities require different levels of attainment of the skill measured (e.g., carpenter versus auto mechanic for an objective dealing with computing the size of various geometric objects). The next basic step is to identify those items that discriminate best between the groups in the desired direction (that is, the presumably more able should do better). It is important for the purposes of CRT interpretation that if two separate groups are involved, they have the same general intellectual ability or other characteristics that might bias the test results.

2. Single Group, Pre- and Posttest. Give the test to the same group twice, once before instruction and again after instruction. Identify those items that discriminate between the two test sessions. A number of item analysis techniques designed specifically for CRTs have used this approach (Popham, 1970; Ozenne, 1971; Kosecoff & Klein, 1973; Roudabush, 1973).

3. Single Group, Posttest Only. Give the test to one group of individuals after a fixed period of instruction, that is, all examinees have had the same amount of opportunity to achieve the objective. If the time allotted is somewhat less than that needed for *all* the students to achieve the objective and the students are somewhat heterogeneous in their ability as is common in most classrooms, then the typical item analysis procedures such as computing point biserial correlation coefficients may be employed to identify faulty items. An internal criterion (total score on test) or an external criterion (success in achieving a more advanced skill) may be used (Glaser, 1963). One weakness in this approach is that items having very high or low difficulties will tend to have low biserial coefficients even though they may be very sensitive to instruction. An extreme case would be an item that would be failed by everyone prior to instruction but passed by everyone after instruction. A second weakness is that general intellectual ability as well as the effects of instruction may influence the results and there is no way of cleanly separating these influences.

4. Single Group, Repeated Measures. Each student periodically takes the complete test until he is able to achieve mastery. A record is kept of the number of times the student passes and fails each item. Analysis is then made to determine whether the item generally exhibits the desired pattern of failure then success (with no reversals), i.e., a desired pattern would be FFPP and an undesirable pattern would be PFFP. This approach is only applicable where there are no carry-over effects from test session to test session or where truly parallel items may be constructed for each test session and then systematically counterbalanced across sessions and examinees. The advantages of this approach are that it permits relevant scaling of an item within an objective and the analysis is made after all students have become "masters." The labor involved in this approach and the likelihood of finding items that scale well, however, have not contributed to this method's popularity.

One issue that is related to item analysis procedures and that seems to be neglected with respect to CRTs is the problem of knowing whether the final set of items provides adequate coverage of the objective. In other words, how many items are really needed to sample sufficiently a given objective? Further, a procedure is needed for determining whether some of the items are redundant. Although these kinds of issues have been examined in part with the more traditional kinds of tests, the unique demands of CRTs will correspondingly require new ways of dealing with this general problem of knowing when one has appropriate and efficient coverage.

Content Validity

A major concern of CRT developers is in establishing the content validity of their instruments. The three most common ways that have been used to do this are as follows:

1. Systematic Test Development. This approach involves presenting the rationale for the systematic procedure employed in terms of why it should result in a content valid test (Hively, 1970, 1973).

2. Expert Judgment. Content experts are given a variety of objectives and the items used to measure them. They are then asked to assign the items to their "appropriate" objective. The degree to which they are able to do this accurately reflects on item-objective consistency and thereby on content validity; that is, is a given item really measuring the objective for which it has been constructed? (Dahl, 1971).

3. Item Analysis. It is possible to compute internal consistency indices for a CRT and/or see whether an item on a given objective correlates more highly with other items for

this objective than it does with items on other objectives. These approaches are limited by all the dangers of internal consistency validation techniques plus the potential problem of no variance on the measures (that is, the students all receive the same score). The latter problem, however, usually appears to be more theoretical than actual, because students do vary in their performance. This variation may be due to a number of factors including the students' general intellectual ability, cultural and environmental backgrounds, and the quality of instruction they receive. If enough students are tested, then one will discover sufficient variance in the levels of performance and/or in the time it takes to achieve a given level. Reports of "no variance" usually stem from failure to sample enough students and/or from the failure to examine the rate at which students master items and objectives. Thus, although one might conceive of a situation in which no variance might occur in a given classroom, it is hard to imagine how this might arise across a variety of classrooms–unless, of course, the test was totally inappropriate for the full range of examinees for whom it was constructed. The real problem, therefore, is not in finding variance but in identifying just that portion of the variance that is due to the student's degree of mastery of the particular objective on which the CRT is based rather than variance due to some extraneous influence.

Item and Test Bias

"Item bias" may be defined as a group by item interaction; that is, the profiles of performance of different groups (such as males versus females) across all items in the test are not parallel. "Test bias" is defined as a group by test interaction; that is, groups do not have the same shaped profile of scores across the various tests being considered (Cleary, 1966; Cleary & Hilton, 1968). Little attention has been paid to CRTs with respect to these kinds of biases, although they have become important topics within the general measurement field.

It should be noted, however, that the identification of a test or set of tests as being "biased" with respect to certain groups does not necessarily mean that the measures should be revised. The reason for this is that such "bias" may only mean that the educational and cultural experiences of the groups taking the tests are systematically different and the basis for these differences and how to deal with them should be examined. It is entirely likely, for example, for a test to appear biased simply because it draws more on the vocabulary from certain texts than it does from others, and the use of the more test-dependent texts is not random in the population of examinees. Wider use of the more dependent texts would, therefore, remove the supposed "bias" in the test; changing the test to be more representative of the texts used would also achieve the same result.

Test Scores

As noted in the preface to this paper, one of the two essential features of a CRT is that an individual's or a group's score on it is interpreted in terms of the level of performance obtained with respect to the achievement of the objective(s) on which the CRT is based. This type of score reporting is contrasted to the norm referenced approach in which a student's or a group's score is interpreted with respect to the performance of other individuals or groups (Popham & Husek, 1969). The primary advantage of the CRT approach is, therefore, its ability to provide a means for describing what the student (or group) can do or what it knows or how it feels without having to consider the skills, knowledge, or attitudes of others.

There is some question, however, as to whether a CRT can really do this (Klein, 1970; Davis, 1971; Ebel, 1972). For example, if parents are told that their child has mastered a given objective or set of objectives, their first question is "Is this performance satisfactory?" In other words, they are asking whether the child is progressing satisfactorily and the only frame of reference one can give in this situation is the rate of progress of other students. The fact that such a normative frame of reference can easily be provided also points out that one can make norm referenced interpretations of CRT scores. The distinctive feature of a CRT score must, therefore, lie in its *emphasis* on describing the absolute rather than the relative level of performance with respect to an objective or skill. Because of this emphasis, different kinds of scores are generally reported for CRTs than for norm referenced measures. Some of the different kinds of scores that can be reported for CRTs that reflect emphasis on objectives are listed below:

1. The number or percent correct on a given objective or set of items than encompass a few highly related objectives.

2. "Mastery" of a given objective or set of items where "mastery" is defined in terms of a certain level of performance such as 90 percent correct.

3. The time it takes (such as in class hours or calendar days) for an individual to achieve a given performance level (including what nas been defined as "mastery" (Harris, 1973).

4. The time (in minutes or hours) it takes a student to perform a certain task or set of tasks related to an objective (such as correctly computing the product of all single digit numerals).

5. The probability that the student is ready to begin the next level of instruction (this may be based on both the number of items correct and the pattern of answers given to these items).

6. The percentage of students who "pass" each item; that

is, the item's difficulty. This kind of score is used exclusively in program evaluation where each item or task is considered important in itself.

Of all the scores listed above, the ones that have been the focus of most discussion are those that imply that the student has achieved "mastery" (Millman, 1972). The reason for this attention is that while such a score comes closest to the underlying spirit of a CRT, there is rarely a good way of defining exactly what is meant by "mastery." Arbitrary definitions, such as 85 percent correct, are rampant; but there is rarely any satisfactory criterion for setting such standards of performance. Further, a mastery score often hides the true level of student performance. In other words, if the student failed to achieve mastery did he miss by a little or miss by a great deal; or if he made it, did he just squeak by? Finally, the problems inherent in the construction of items for a CRT and especially those dealing with the defining of the acceptable item types, item selection procedures, and item difficulty severely limit the interpretation of what is meant by "mastery."

Packaging and Other Considerations

How a CRT is finally put together and packaged is again a function of the purpose(s) for which it will be used relative to the various kinds of constraints imposed on its development and use. When there is a vast number of objectives to be assessed and it is not considered reasonable to develop a separate CRT for each, one or more of the following techniques are used:

1. Combine objectives that are considered highly related to one another into a single measure.
2. Select a group of objectives from the total pool of objectives based on a set of appropriate criteria (such as those presented in the Appendix).
3. Limit the scope of each objective so as to reduce the potential number of items and/or tasks that might be needed to measure it.

All of these techniques do, of course, require the use of experts in the fields of measurement and curriculum in order to make sound compromises from both content and methodological points of view.

The methods of packaging and distributing CRTs are quite varied. One of the potentially most functional techniques involves printing tests on spirit masters so that each teacher can duplicate the copies needed for a given class without having to purchase large numbers of test booklets. A second innovation that appears to have promise is referencing the objective and even the test item to specific instructional materials. In one such case, the test form was printed in such a way that the teacher was told immediately whether the student passed the item, and in the event of a failure a manual then directed teachers to materials for additional instruction.

PRESENT EFFORTS IN CRITERION REFERENCED TESTING

This section of the paper summarizes the results of a survey conducted to assess current efforts in criterion referenced testing. All information is based on data provided directly by the projects themselves or through associated technical reports, journal articles, and interviews.

Although special emphasis was given to criterion referenced measures in mathematics, related developmental efforts in other content areas were also reviewed. The list of projects reported here is not exhaustive[3] but can be viewed as representative of the general state of the art in criterion referenced testing.

Five defining characteristics of criterion referenced testing programs have been identified. They include program focus, instructional dependence, [4] objective and item generation, test models and packaging, and test scores. Each of these characteristics has already been discussed in the first section of this paper, however, some further explanation regarding the scale used for the instructional dependence category is needed.

California Test Bureau - McGraw-Hill (CTB) Prescriptive Mathematics Inventory (PMI)

Focus. CTB is interested in the construction of CRT programs for classroom management. In particular the PMI was designed to measure 351 objectives representing the mathematics curriculum nominally taught in grades four through eight.

[3] The projects reported in this section are those that responded to our survey. Projects were selected for the survey on the basis of an extensive ERIC search and general knowledge of the field. It can be expected, therefore, that some CR testing efforts may have been overlooked or that some programs did not respond.

[4] A dichotomous classification is used to describe a criterion referenced program's degree of instructional bias. Programs with a large degree of instructional dependence develop test items that are dependent on a particular curriculum or set of instructional materials and techniques. Programs with a small degree of instructional dependence, on the other hand, construct test items that are not dependent on the specific skills or content of an instructional program.

Instructional Dependency: Small. Neither the objectives nor the test items reflect any instructional bias.

Objective and Item Generation. Using a "consensus approach" objectives were culled from the text materials most widely used in schools, collated from each source into a single list, classified into broader objectives classifications, and analyzed with respect to content and a hierarchical structure. Items were then developed to measure these objectives. (Note: On the PMI only one item is used to assess each objective.)

Test Model and Packaging. The PMI is divided into four levels based on the objectives most commonly taught in grades 4 and 5, 5 and 6, 6 and 7, and 7 and 8. The test items sample various levels of difficulty in each of the content categories represented. In responding to the PMI, the student records his answers on unique, item specific machine-scoreable answer grids specially designed to eliminate guessing.

In addition to the actual PMI test, CTB/McGraw-Hill offers the following support materials and services:

- Complete scoring and reporting services (that provide information on objectives mastered and not yet mastered)
- Practice exercise booklets, an examiner's manual and class information sheet (to identify the class and tests)
- An Individual Diagnostic Matrix (reporting the student's score on each objective)
- A Class Diagnostic Matrix (reporting average class scores on each objective)
- An Individual Study Guide (that references pages in texts where material can be found for objectives which the student did not master)
- A Class Grouping Report (that lists students according to their deficiencies in major content areas)

Test Scores. Because one item is used to measure each behavior, the mastery criterion for each objective is that the student correctly solve the associated item (Roudabush, 1971). Test scores are then reported in terms of mastery or non-mastery for each objective.

Four different types of reports are available for reporting test scores: two individual reports for each student, and two reports for the class. The Individual Diagnostic Matrix shows a profile of the student's mastery or non-mastery of the objectives. The Individual Study Guide gives page references for a selected textbook covering those objectives not yet mastered by the student. The Class Diagnostic Matrix summarizes test results for the whole class in terms of the percentage of students mastering each objective. And finally, the Class Grouping Report indicates how students fall into achievement groups within the mathematics curriculum and provides page references to the textbook being used in the classroom for materials covering objectives that were frequently missed.

Additional information available from:

CTB/McGraw-Hill
Del Monte Research Park
Monterey, California 93940

Comprehensive Achievement Monitoring (CAM)

Focus. CAM is designed as a computer-assisted, multipurpose evaluation system useful at individual, group, district, or state levels.

The CAM model is based on two attributes: (1) a flexible time series design (testing at frequent intervals) which can be varied to meet the financial limitations and information needs of the user, and (2) a procedure for sampling students and items which both introduces economy into testing and increases the comprehensiveness of behavior samples available from each testing session.

At present, the New York State Department of Education has installed CAM or modifications of CAM in five school districts. Although programs have mostly involved math, they are currently being extended to science and reading.

Instructional Dependency: Large. CAM is constructed to be most effective when the items relate directly to course objectives.

Objective and Item Generation. Curricula are defined by behavioral objectives which are systematically coded for easy identification, retrieval, and grouping, and by one or more classifications. This process is typically carried out by potential system users (that is, teacher groups).

With respect to objectives specification, a "behavioral analysis" of course content requires that the user (1) prepare a topical course outline, (2) specify the general course objectives derived from the content (in non-behavioral terms), (3) specify the terminal course objectives (in behavioral terms), and (4) specify enabling objectives (in behavioral terms). Objectives are then organized into classifications, typically utilizing Ammerman and Melching's (1966) classification system for the specificity of instructional objectives by their relationship to terminal student performance.

Items are developed by system users (teachers) directly from objectives and are then judged (typically by the item writers themselves) for their consistency with the objectives. Considerations of error from guessing, ease of scoring, criterion referenced versus norm referenced test interpretations, and general item writing skills (that is, "the item stem must be worded to require specific response") guide item construction activities.

Test Model and Packaging. The typical set of CAM tests is constructed around the stated objectives of the course or program to be evaluated. Objectives, items, and test forms are typically generated by system users in accordance with instructions provided in a user's manual.

Generally, a pool of items is constructed with approximately 4 to 10 samples per objective. Through random stratified sampling items are assigned to test forms creating parallel test forms or monitors. Students receive the test forms in a random order at fixed testing intervals (Determined by the user's information needs). Each test form contains a fixed number of items representing objectives which are taught between test administrations. Test forms are usually short, requiring from 10 to 30 minutes of testing time.

Test Scores. Through sampling of test items and testing at frequent intervals, CAM generates performance data on all course objectives in relation to three phases of time: before instruction, immediately after instruction, and retention over long periods of time.

After each test administration each student receives a report listing the correct and incorrect responses to every item as well as total scores on current and previous tests. Group data are also provided in the form of percent achievement by designated objectives for each test administration. Finally, achievement profiles which graphically display the level of achievement (in terms of percent correct test scores) for all previous and current tests on selected objectives are available quarterly.

Additional information available from:

Robert O'Reilly
Chief, Bureau of School & Cultural Research
University of the State of New York
State Education Department
Albany, New York 12224

William Gorth
School of Education
University of Massachusetts
Amherst, Massachusetts 01002

Individualized Criterion Referenced Testing (ICRT)

Focus. ICRT offers criterion referenced testing programs emphasizing individual student achievement and providing two basic kinds of information: first, the specific knowledge and skills which the student has learned, and second, the specific knowledge and skills which are the next instructional steps to be mastered. At present such testing programs are available in reading and mathematics; the following comments will focus primarily on the mathematics system.

Instructional Dependency: Large. The basis for the criterion referenced tests is a set of specified instructional objectives which describe the Continuous Progress Laboratory's Math program.

Objective and Item Generation: Instructional objectives referenced to the Continuous Progress Laboratory's math curriculum are arranged from the most elementary to the most difficult, forming an instructional continuum. From this instructional continuum those objectives common to most curricula and expected of most students are selected as testing objectives. These selected objectives, arranged with respect to item difficulty, constitute a testing continuum. The testing continuum is then used as a basis for item and test generation.

Test Models and Packaging: ICRT provides test kits for each grade level 1-8. Each test kit has sufficient tests for an average class, a Teacher's Manual, a scoring template and an orientation kit. In addition, each kit (with the exception of level 1) contains multiple copies of the grade level test booklets as well as multiple copies of booklets for up to two levels below the indicated grade level of the kit.

Tests are designed to be self-administered or administered with teacher guidance. All the tests are power tests with no implied time limit. Each test has approximately 16 items (2 items per objective). The student records his responses to the test items on computer cards. Directions for test scoring are included in the teacher's guide.

Four kinds of score reports are available: a District Summary, a Building Summary, a Class Summary, and a Student Summary. The Student Summary provides prescriptive instructional resources.

Test Scores: Students' scores on each objective are reported to District, Building, and Class Summaries; students' scores are reported in terms of how many students are at various working levels (a student's approximate working level is determined by the first test booklet in which he or she missed 3 or more objectives). The Student Summary is intended as a prescriptive instrument, indicating which objectives have been mastered, which require review, and which should be learned next. In addition, prescriptive instructional resources are suggested for objectives which the student needs to review or learn. These prescriptive guides are referenced to the Continuous Progress Laboratory Math Program, the supplementary drill tapes, and three additional curricula selected by the user.

Additional information available from:

Louis Miller, Vice President
Educational Progress Corporation
3000 Sand Hill Road
Menlo Park, California 94025

Charles Carlson
Educational Progress Corporation
4900 South Lewis Avenue
Tulsa, Oklahoma 74105

Instructional Objectives Exchange (IOX)

Focus. A criterion referenced test program has been developed to complement the IOX objectives collections. The decision to develop these objectives based tests represents an effort to provide readily usable support materials to assess individual student progress and to facilitate classroom management.

Instructional Dependency: Small. Neither the objectives booklets nor the criterion referenced tests are based on any particular curriculum or instructional program.

Objective and Item Generation. Within each subject area objectives are defined in terms of relevant topics and skills at three levels of generality. Criteria for sampling the most general categories include importance of the area, economy of production into tests, and practical scoreability. Selection of the type of learner behavior to serve as the specific objective is then guided by considerations of transferability or generalizability within a content area, importance, terminality (that is, the highest step in a hierarchy), transferability outside the area, ease in scoring, and amenability to instruction.

Rooted in Wells Hively's (1970, 1973) item form analysis, expanded objectives (called amplified objectives) are used to define permissible stimulus and response options for item generation. For each objective only one type of test item is used; the associated item format is then carefully defined by an amplified objective.

Test Models and Packaging. IOX provides manuals listing objectives, sets of criterion referenced tests, and a user's guide or test manual. In the area of elementary mathematics, for example, there are five independent sets of criterion referenced measures which cover the nine mathematics strands identified by the California State Department of Education. For each set of tests a parallel set is available to facilitate pre- and posttesting (that is, each set of tests is available in a form A and a form B which contain parallel tests).

Tests are distributed on one page, preprinted spirit masters which can be used by teachers to duplicate sufficient copies for their students. The typical test is multiple choice in format, contains 5 to 10 items and requires about 30 minutes to complete. The test manual provides a list of objectives in that area, sample test items, complete instructions for test administration, answer keys, and a guide for classifying scores in terms of achievement levels (whether or not the student attained mastery).

Test Scores. Although directions are provided in a user's guide for classifying scores into mastery groups, IOX does not provide forms for reporting scores or suggestions for tabulating test scores.

Additional information available from:

Instructional Objectives Exchange
Box 24095
Los Angeles, California 90024

MINNEMAST Curriculum Project - University of Minnesota

Focus. The MINNEMAST Project represents an experimental effort to develop a coordinated and sequential mathematics and science curriculum for the elementary school. As part of the evaluation of this project, a technology for criterion referenced test construction was developed by Hively and his associates at the University of Minnesota. These tests were primarily intended to assess the MINNEMAST Program itself rather than individual students' progress.

Instructional Dependency: Small. Test items were generated that reflect the entire range of skills and behaviors associated with a given objective.

Objective and Item Generation. Relevant learner behaviors and skills associated with a given content area were organized (by the MINNEMAST staff) into classes called learning domains. The basic notion underlying this process is that important classes of content and skill would be completely defined in terms of behaviorally stated, structured sets or domains.

Rules for generating test items for a given learning domain are organized into formal schemes called item forms. There are three major components to an item form: (1) instructions (directions given to students), (2) stimulus characteristics (the skills and behaviors an item can cover and rules for constructing specific kinds of items), and (3) response characteristics (acceptable way of responding to an item, for example, written or oral responses).

Test Models and Packaging. (It should be noted that the MINNEMAST efforts reported here were field test activities, and consequently a final packaging mode was not available).

The MINNEMAST curriculum was divided into discrete units. For each unit the teacher was provided with a handbook containing a sequence of lessons, general statements about goals, explanatory background information, and lists of materials needed for lessons.

Test construction was computer-assisted and conducted by the MINNEMAST staff. A system of student-item sampling was utilized to gather information on all test items with a minimum of testing time. To this end computer printout labels were generated for each student listing his or her name, identifying data such as class and school, and the items assigned to him or her. When all the items specified from an item form had been written the computer printout

labels were attached to them and the items were then collated into tests for the individual students.

Test Scores. The principal data derived from the MINNEMAST testing program were the proportion of correct responses. Whenever possible, however, additional information was reported concerning the kinds of correct and incorrect responses being made.

Although no set format for reporting scores was stipulated, data were usually presented in tables showing complete item-by-item listings of actual responses as well as frequencies of various categories of responses (for example, frequencies for individual items, item forms or objectives, and groups of objectives). Due to the absence of empirical evidence, desired levels of achievement were not established in advance of testing.

Additional information available from:

Wells Hively
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

National Assessment of Educational Progress (NAEP)

Focus. The purpose of NAEP is the assessment of educational attainments on a national basis.

Instructional Dependency: Small. Neither objectives nor items are referenced to any curriculum text or instructional program.

Objective and Item Generation. NAEP defines its objectives and the associated skills and behaviors (the "domain of reference") through a national consensus of opinion regarding the important goals and outcomes of education with respect to a given subject area.

Objectives developed by NAEP's Exercise Development Department are reviewed by external subject matter experts and layman groups. Following the development of objectives, contracts are awarded for item generation. The amount of items developed for a given objective is based on a weighting scheme determined by the subject matter experts. A framework for item writing is provided by a system of exercise prototypes that define four characteristics of an item: (1) administrative mode (can the item be administered individually or to a group), (2) stimulus mode (audio, visual, and so on), (3) response mode (multiple choice or free response), and (4) response category (written, verbal, role playing, and so on).

Test Models and Packaging. Tests are designed exclusively for measuring student achievement on a national scale. The number of items for a given objective is determined by a weighting scale based on priorities identified by the subject matter experts. Tests are available at four age levels (9, 13, 17, and adult). Two subject areas are currently being assessed each year with a five-year reassessment cycle. (Mathematics is scheduled for the 72-73 school year.) Two hundred and ten minutes of testing are allotted annually to each subject at each age level.

Test Scores. Scores are generally reported as the percentage of correct responses by items and for various classes of items. For example, items dealing with solving algebraic equations might be compared with items on mathematical induction. In addition, scores are broken down in terms of typical performance by region, sex, SES, and so on.

Additional information available from:

National Assessment of Educational Progress
822 Lincoln Tower
1860 Lincoln Street
Denver, Colorado 80203

Southwest Regional Laboratory (SWRL)

Focus. SWRL is involved in the development of text-referenced instructional management systems that operate in conjunction with a developed curriculum. At present such a classroom management system in reading is available at the kindergarten level and a math system is under development. Criterion referenced tests have been incorporated into this system to assess student progress.

Instructional Dependency: Large. The SWRL program is specifically based on a predefined curriculum: "... to be minimally useful the (CR) test must be specifically referenced to a prespecified structure of achievement. To be maximally useful the tests must be specifically referenced to defined instructional materials" (Baker, R.L., 1972).

Objective and Item Generation. Hively's (1973) item form approach and related processes are utilized to define classes of behaviors and skills associated with specific content areas. A collection of item forms sequentially organized together with a list of constraints on item generation provide the framework for defining total content areas in behavioral terms (a "universe of content"). Strings of item forms are then organized into tentative sequences or "instructional specifications" that map out the instructional and evaluation efforts consistent with the item forms.

Test Models and Packaging. With respect to evaluation activities each instructional management system provides:

- A means (vis-a-vis testing) for student placement
- Criterion referenced measures on 3 to 8 instructional

outcomes 10 to 15 times during the year. (Note: These tests are constructed for specific information purposes—to assess student progress on objectives attended to by a specific curriculum.)

- Additional practice materials for the instructional outcomes which have continuity throughout the text
- A mid-year and end-of-year evaluation measure
- A Quality Assurance System (a user's manual providing directions and pacing information)

Test Scores. The Quality Assurance Manual provides forms for reporting the means, standard deviations, and percent of students attaining criterion performance. Regression analyses between criterion scores on final and mid-year criterion referenced tests are also reported based on a large student sample.

Additional information available from:

Southwest Regional Laboratory for
Educational Research and Development
4665 Lampson Avenue
Los Alamitos, California 90720

System for Objectives Based Assessment—Reading (SOBAR)
Center for the Study of Evaluation: UCLA

Focus. SOBAR basically constitutes an item bank integrated into a selection/delivery system intended as a multipurpose evaluation procedure appropriate at the individual, group, or program level. Designed to serve as an exemplary objectives based assessment system, SOBAR includes a set of performance objectives covering the entire spectrum of a content area (in this case that of reading, grades K-12), a classification system for selecting objectives, and a bank of assessment items keyed to specific objectives.

Instructional Dependency: Small. SOBAR is seen as a flexible, multipurpose test generation system that is not dependent on a given instructional program or information need.

Objective and Item Generation. A set of objectives was developed by the SOBAR staff (with the help of reading experts) to cover the complete content area of reading. These objectives were then classified into categories reflecting various skill areas and levels of generality. Upon completion of objective specification the SOBAR staff constructed items keyed to the objectives. During item writing special attention was given to the independence (nonredundancy) of items, objective item congruence, and the comprehensiveness of items. The system is referenced to performance objectives.

Test Models and Packaging. Among the materials and services provided to users SOBAR includes:

- A comprehensive catalogue of nearly 500 objectives. (These objectives cover grades K-12 and are divided into six major skill categories.)
- A guide and selection chart to aid the user in selecting objectives appropriate to local priorities
- Computer generated reports of the outcome of the objective selection process
- Tests for each SOBAR objective. These measures are leveled by grade clusters: K-3, 4-6, 7-9, and 10-12. Depending on the nature of the objective a test for an individual objective may contain 5-20 test items.

Test construction is viewed in terms of the user's specific information needs. Items are selected for tests according to the test model appropriate for a given test situation. In addition, tests can be assembled at different levels of objective generality.

Test Scores. At present SOBAR has not begun to field test methods of score reporting and interpretation.

Additional information available from:

SOBAR Project
Center for the Study of Evaluation
University of California
145 Moore Hall
Los Angeles, California 90024

Zweig and Associates

Focus. Zweig offers a criterion referenced testing program based on behavioral objectives and indexed to prescriptions for teaching alternatives. At present such testing programs designed for classroom management within the context of individualized instruction are available in reading and mathematics. The Fountain Valley Teacher Support System in mathematics was reviewed for this paper. Comments are largely based on the Fountain Valley System.

Instructional Dependency: Small. Objectives and items cover the entire spectrum of skills reflected in the nine mathematics strands for California.

Objective and Item Generation. Objectives and items are generated by teacher groups (followed by a review from experts) and reflect skills in each of the nine mathematical strands developed by the California State Department of Education. Strands are measured at each grade level, K-8, for which there is pertinent instruction. Typically 3 to 5 items are constructed for each objective.

Test Models and Packaging. The Fountain Valley System includes:

- 785 objectives organized by strand and grade level
- 196 self-scoring, self-administering tests
- Continuous Pupil Progress Profiles to record individual student achievement
- Class ditto masters to document group performance
- Teacher Manuals for each grade level (that include a listing of all objectives at that level)
- Manuals of criterion referenced teaching alternatives

All materials are color coded. Tape cassettes at each level provide directions for test administration. Each test is printed on a sealed form made of treated paper that automatically records student's responses on the reverse side of the test sheet as the student takes the test. In addition, the reverse side indicates correct or incorrect responses by a number code which corresponds to the objective and strand being tested and provides a score interpretation key to classify scores into "proceed" and "reteach" categories.

At each grade level a Teaching Alternatives Manual documents (by number code) prescriptive activities (for skills falling into the reteach category) listed by number code under each publisher's name and series.

Test Scores. Student scores in each skill for each of the nine strands are recorded on a Continuous Pupil Progress Profile (CPPP). The objectives for each strand are arranged on the CPPP in a hierarchy of difficulty, grouped by grade levels, and designated by color and number codes. Objectives measured by each test are then grouped between heavy lines. Student scores are recorded on the CPPP as either reteach or proceed in accordance with scoring instructions on the answer sheet. These instructions give the number of incorrect answers that determine the classification for each skill.

Additional information available from:

Richard L. Zweig Associates, Inc.
20800 Beach Boulevard
Huntington Beach, California 92648

Summary and Conclusions

This paper has attempted to outline the basic steps and procedures in the development of criterion referenced tests as well as the issues and problems associated with these activities. In addition, representative CRT systems have been reviewed. From this analysis it is clear that the developer of a CRT must answer a number of questions in order to clarify the nature and purpose of a CRT.

1. For what decision areas and purposes is the CRT most applicable?
2. What areas and objectives does the CRT cover and how were these objectives derived and organized?
3. How broadly or narrowly are the objectives defined?
4. How were the test items or tasks chosen to measure the objectives defined and developed?
5. How dependent are the items on particular instructional materials or programs? And what is their applicability to different kinds of students?
6. What methods were used to improve the items on the CRT and why were they chosen relative to the purpose of the instrument?
7. How was the validity of the CRT established?
8. What kinds of scores should be reported for a CRT and what is the justification for these scores, especially those involving "mastery?"
9. How was the test finally put together, what compromises had to be made, and how were they resolved?
10. In what ways will packaging of the CRT facilitate its use?

These questions will hopefully serve three functions. The first is that they will guide CRT developers to the issues that must be addressed in both the construction process and in the manual that accompanies the final instrument. The second purpose these questions may serve is to guide researchers to those problems of major interest within the field of criterion referenced testing. Finally, they will help the purchasers of CRTs to understand better the kinds of variables they must consider in order to make a wise selection of instruments and an appropriate interpretation of the results obtained with them. Certainly the publication of a set of minimum standards for CRTs by an appropriate professional organization would go a long way toward ensuring that these functions have been carried out successfully.

APPENDIX: IOX Criteria for Selecting Objectives*

The following criteria should be applied in deciding on the type of learner behavior which will serve as the specific objective, thereafter to guide the test construction:

(1) *Transferability Within Domain*. The form of learner behavior selected should be the most generalizable of those represented in the content general domain, i.e., a learner mastering the designated behavior requirements would likely be able to transfer that mastery to most, if not all, of the other eligible behavioral requirements in the content general domain.

In making such a selection it is important to consider the entire range of learner behaviors with which we are concerned, i.e., both test-like events and real world events. For instance, in surveying an individual's mathematical competence we should be attentive not only to the X_1, X_2, and X_3, which we can represent via standard test formats but to the X_{17}, X_{18}, and X_{19}, which might reflect such skills as the ability to make change in a supermarket or to complete one's annual income tax report.

The test constructor should sketch out as wide a range of alternatives as possible, then select the one testable learner behavior which will most readily transfer to the other learner behaviors delimited by the content general objective.

(2) *Widely Accepted*. The objective selected should be the most widely accepted as important by those in the field. Unlike the IOX objective collections where we present a wide array of alternatives and then encourage educators to choose among them, here we will have to go with the majority preference. Clearly, this criterion is not unrelated to criterion number one, but it may be profitable to apply it independently.

(3) *Terminality*. If there is a degree of possible hierarchy present in the contending types of learner behaviors under consideration, such that some are considered precursive or enroute to others, the chosen specific objective should represent the most terminal learner behavior.

(4) *Transferability Outside the Domain*. Another consideration in selecting a specific objective is the degree to which that behavior, once mastered, will be transferable outside the content general domain, for example, to domains which might be learned by students in the future. For instance, certain skills acquired by students in one course (such as the ability to distinguish between fact and opinion) may have reference to many other courses. Such high transfer skills and intellectual constructs should be given high priority in the selection of specific objectives.

(5) *Ease of Scorability*. In an effort to produce tests which have considerable practical utility, we must try to select learner behaviors which, other factors being equal, can be easily scored by those educators employing them. Again, this does not limit us to selected response items, for in some instances we shall surely find it necessary to utilize constructed response formats. (This may help distinguish the IOX tests from typical standardized tests.) Nevertheless, scoring practicality is a nontrivial consideration.

Now how should these five criteria be employed in selecting the specific objectives? Should they be weighted equally, in descending order, or in reverse order (stratified according to number of two syllable words in the descriptive paragraphs)? Sorry, but no handy scheme is available for mechanical translation into decisions. Test constructors must, however, be self-consciously attentive to each of the five points. We may devise a check sheet or other shorthand form to encourage such attention. If the test developer has exhausted all rational alternatives, an arbitrary selection is always possible.

Having chosen the specific objectives, that is, the categories to be used in generating a pool of homogeneous test items which assess a given learner behavior, the next task involves the production of a defensible set of such items.

---

*Excerpted with permission of the author, W.J. Popham, from *Selecting Objectives and Generating Test Items for Objectives-Based Tests*, Los Angeles, IOX, 1972.

REFERENCES*

Airasian, P., & Madaus, G. Criterion referenced testing in the classroom. *Measurement in Education*, 1972, 3 (4), 1-8.

Baker, E.L. Using measurement to improve instruction. Paper presented at Convention of American Psychological Association, Honolulu, Hawaii, 1972. ED 069 762.

Baker, R.L. Measurement considerations in instruction product development. Paper presented at Conference on Problems in Objectives Based Measurement, Center for the Study of Evaluation, University of California, 1972.

Bormuth, J.P. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

Cleary, T. Test bias: Validity of the scholastic aptitude test for negro and white students in integrated colleges. Research Bulletin 66-31. Princeton, New Jersey: Educational Testing Service, 1966. ED 018 200.

Cleary, T., & Hilton, T. An investigation of item bias. *Educational and Psychological Measurement*, 1968, 28 (1), 61-75.

Cox, R., & Vargus, J.C. A comparision of item selection techniques for norm referenced and criterion referenced tests. Pittsburgh: Center for the Study of Instructional Programs, Learning Research and Development Center, University of Pittsburgh, 1966.

Cronbach, L. Test validation. In L. Thorndike (Ed.), *Educational Measurement* (2nd ed). Washington, D.C.: American Council on Education, 1971.

Dahl, T.A. The measurement of congruence between learning objectives and test items. Unpublished doctoral dissertation, University of California, Los Angeles, 1971.

Davis, F.B. Criterion referenced tests. Paper presented at Annual AERA Meeting, New York, 1971. ED 050 154.

Davis, F.B. Criterion referenced measurement. 1971 AERA Conference Summaries, ERIC/TM Report 12, 1972. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1972. ED 060 134.

Davis, F.B. Criterion referenced measurement. 1972 AERA Conference Summaries, ERIC/TM Report 17, 1973. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1973. ED 073 143.

Ebel, R.L. Evaluation and educational objectives: Behavioral and otherwise. Paper presented at the Convention of the American Psychological Association, Honolulu, Hawaii, 1972.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 1963, 18, 519-521.

Glaser, R., & Nitko, A. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971. Pp. 652-670.

Harris, C. Comments on problems of objectives based measurement. Paper presented at Annual AERA meeting, New Orleans, 1973.

Hively, W. Introduction to domain referenced achievement testing. Symposium presentation, AERA, Minnesota, 1970.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project. CSE Monograph Series in Evaluation, Volume 1. Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Keller, C.M. Criterion referenced measurement: A bibliography. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1972. ED 060 041. . . bibliography ERIC/TM Report 7, 1972.

Klein, S.P. Evaluating tests in terms of the information they provide. *Evaluation: Comment*, 1970, 2 (2), 1-6. ED 045 699.

Klein, S.P. An evaluation of New Mexico's educational priorities. Paper presented at Western Psychological Association, Portland, 1972. TM 002 735. (ED number not yet available.)

Kosecoff, J.B. & Klein, S.P. Analyzing tests and test items for sensitivity to instructional effects. CSE Working Paper No. 24, Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Kriewall, T.E., & Hirsch, E. The development and interpretation of criterion referenced tests. Paper presented at Annual AERA Meeting, Los Angeles, California, 1969. ED 042 815.

Mager, R. F. *Preparing instructional objectives*. San Francisco: Fearon Publishers, Inc., 1962.

Millman, J. Passing scores and test lengths for domain referenced measures. Paper presented at Annual AERA Meeting, Chicago, 1972. ED 065 555.

*Items followed by an ED number (for example ED 069 762) are available from the ERIC Document Reproduction Service (EDRS). Consult the most recent issue of *Research in Education* for the address and ordering information.

Nitko, A.J. A model for criterion referenced tests based on use. Paper presented at Annual AERA Meeting, New York, 1971. ED 049 318.

Nitko, A.J. Problems in the development of criterion referenced tests. Paper presented at Annual AERA Meeting, New Orleans, 1973.

Ozenne, D.O. Toward an evaluative methodology for criterion referenced measures: Test sensitivity. CSE Report 72, Center for the Study of Evaluation, University of California, Los Angeles, 1971. ED 061 263.

Popham, W.J. *The teacher-empiricist; A curriculum and instruction supplement*. Los Angeles: Lennox-Brown, Inc., 1965.

Popham, W., & Husek, T.R. Implications of criterion referenced measurement. *Journal of Educational Measurement*, 1967, 6 (1), 1-9.

Popham, W. Indices of adequacy for criterion referenced test items. Presentation at Joint Session of NCEM and AERA, Minneapolis, Minnesota, 1970.

Popham, W.J. Selecting objectives and generating test items for objectives based tests. Paper presented at Conference on Problems in Objectives Based Measurement, Center for the Study of Evaluation, University of California, Los Angeles, 1972.

Roudabush, G. Some reliability problems in a criterion referenced test. Paper presented at Annual AERA Meeting, New York, 1971. ED 050 144.

Roudabush, G.E. Item selection of criterion referenced tests. Paper presented at Annual AERA Meeting, New Orleans, 1973. ED 074 147.

Skager, R. Generating criterion referenced tests from objectives based assessment systems: Unsolved problems in test development, assembly and interpretation. Paper presented at Annual AERA Meeting, New Orleans, 1973.

Wilson, H.A. A humanistic approach to criterion referenced testing. Paper presented at Annual AERA Meeting, New Orleans, 1973.

Zweig, R., & Associates. Personal communication, March 15, 1973.

Selected References on Test Item Construction

Ebel, Robert L. *Essentials of educational measurement*. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1972.

Gronlund, N.E. *Constructing achievement tests*. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Wood, Dorothy A. *Test construction*. Columbus, Ohio: Merrill, 1961.